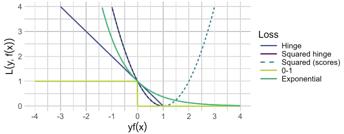
Introduction to Machine Learning

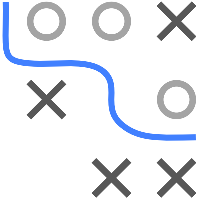
Advanced Risk Minimization
Advanced Classification Losses



Learning goals

- (squared) Hinge loss
- *L*2 loss defined on scores
- Exponential loss
- AUC loss

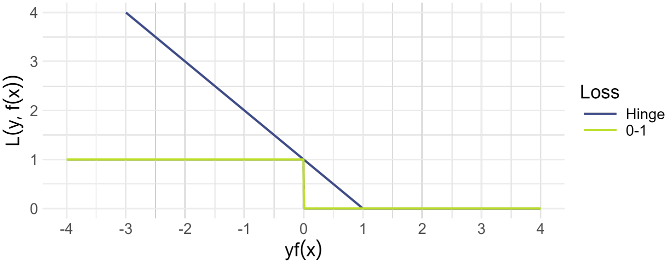
HINGE LOSS



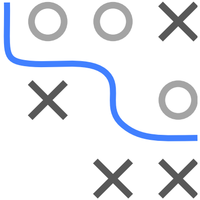
- 0-1-loss intuitive but ill-suited for direct optimization
- **Hinge loss** is continuous and convex upper bound on 0-1-loss

$$L(y, f(\mathbf{x})) = \max\{0, 1 - yf(\mathbf{x})\} \quad \text{for } y \in \{-1, +1\}$$

- Only zero for margin $yf(\mathbf{x}) \geq 1$, encourages confident predictions
- Often used in SVMs
- Resembles a door hinge, hence the name



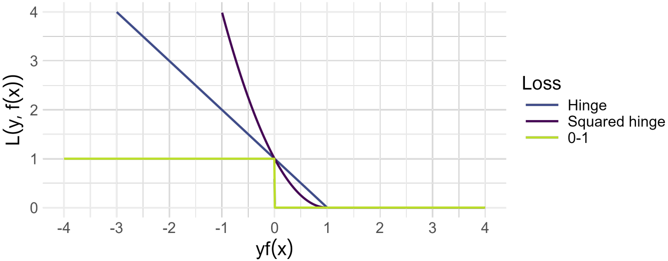
SQUARED HINGE LOSS



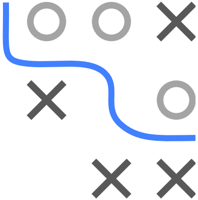
- Can also define **squared hinge loss**:

$$L\left(y, f(\mathbf{x})\right) = \max\{0, (1 - yf(\mathbf{x}))\}^2$$

- $L2$ form punishes margins $yf(\mathbf{x}) \in (0, 1)$ less severely but puts high penalty on confidently wrong predictions
- Cont. differentiable yet more outlier-sensitive than hinge loss



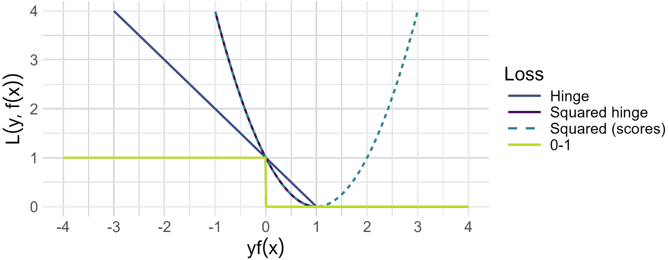
SQUARED LOSS ON SCORES



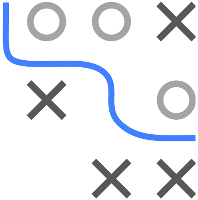
- Analogous to Brier score on probs, can specify **squared loss on classification scores** with $y \in \{-1, +1\}$ using $y^2 = 1$:

$$\begin{aligned} L(y, f(\mathbf{x})) &= (y - f(\mathbf{x}))^2 = y^2 - 2yf(\mathbf{x}) + f(\mathbf{x})^2 \\ &= 1 - 2yf(\mathbf{x}) + (yf(\mathbf{x}))^2 = (1 - yf(\mathbf{x}))^2 \end{aligned}$$

- Like sq. hinge loss for $yf(\mathbf{x}) < 1$, but not clipped to 0 for $yf(\mathbf{x}) > 1$
- Only 0 for $yf(\mathbf{x}) = 1$ and increasing again in $yf(\mathbf{x})$ (undesirable!)



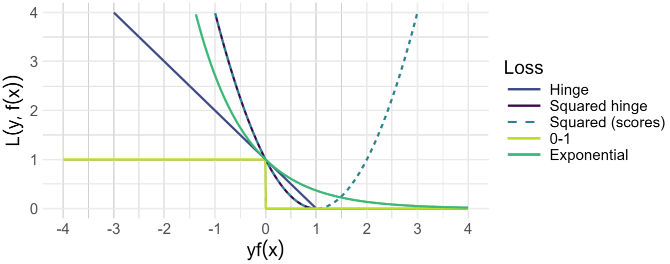
EXPONENTIAL LOSS



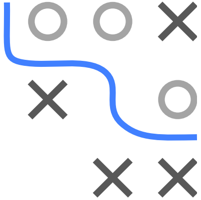
- Another smooth approx. of 0-1-loss is **exponential loss**:

$$L(y, f(\mathbf{x})) = \exp(-yf(\mathbf{x}))$$

- Used in AdaBoost
- Convex, differentiable (thus easier to optimize than 0-1-loss)
- Loss increases exponentially for wrong predictions with high confidence; low-confidence correct predictions have positive loss



AUC-LOSS



- AUC often used as evaluation criterion for binary classifiers
- Let $y \in \{-1, +1\}$ with n_- negative and n_+ positive samples
- AUC can then be defined as

$$AUC = \frac{1}{n_+} \frac{1}{n_-} \sum_{i:y^{(i)}=1} \sum_{j:y^{(j)}=-1} \mathbb{I}[f^{(i)} > f^{(j)}]$$

- Not differentiable w.r.t f due to indicator $\mathbb{I}[f^{(i)} > f^{(j)}]$
- Indicator can be approximated by distribution function of triangular distribution on $[-1, 1]$ with mean 0
- Direct optimization of AUC numerically difficult, rather use common loss and tune for AUC in practice

Comprehensive survey on advanced loss functions: ▸ Wang et al. 2020